Exploring the Relationship Between Bias and Speech Acoustics in Automatic Speech Recognition Systems

An Experimental Investigation Using Acoustic Embeddings and Bias Metrics on a Dataset of Spoken Dutch

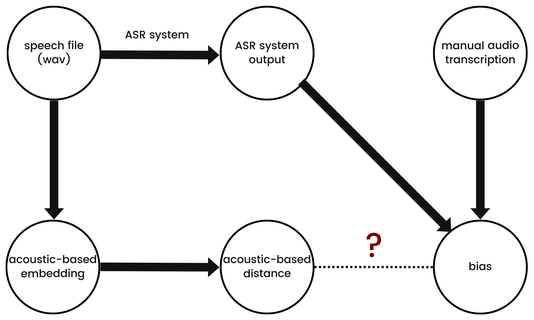
01 Introduction

- Automatic Speech Recognition (ASR) systems convert speech to text.
- These systems were shown to have a bias that manifests itself in recognising the speech of different demographic groups with different accuracy, for example, depending on the speaker's race, gender, or age.
- Bias in ASR systems can have multiple origins, including the quality and diversity of the training dataset and the diversity of the developer team.
- The direct cause of bias lies in how an ASR processes the speech input.
- Previous studies have analyzed the relationship between bias and phonemes, revealing that certain phonemes are more prone to misrecognition, which can contribute to bias against specific groups of speakers.
- In this research, we explore the relationship between bias and acoustic variation of the speakers.

02 Research questions

- How are the bias of an ASR system and the acoustics of the speaker related?
- Which method of capturing acoustic features best reflects the bias?

03 Methodology



- The dataset consists of speech files, ASR output, human-made ground truth transcription.
- In this study, the speech of 100 native Dutch children was used.
- The audio files come in two types of speech: read and human-machine interaction speech (HMI).
- The bias against speaker S is defined as the difference in word error rate between S's speech and the lowest word error rate in the dataset.
- The acoustic embeddings (wav2vec 2.0, XLSR) are in the form of 1024 by X matrix, where X is dependent on the length of the speech.
- The acoustic distance is given by the distance given by Dynamic Time Warping (DTW) known as the sequence alignment algorithm.
- The interrelation metric between the bias and acoustic distance is correlation.

04 Results

ASR system	Read		HMI	
	w2v2	XLSR	w2v2	XLSR
NoAug	0.594	0.550	-0.023	-0.264
SpAug	0.538	0.445	0.191	-0.069
SpSpecAug	0.553	0.500	0.362	0.030
Ws	0.537	0.471	0.029	-0.088
WsFT_cgn	0.473	0.425	-0.052	-0.042

Table 1: Correlation between the bias and acoustic distance for different models with wav2vec 2.0 and XLSR for HMI and Read speech types.

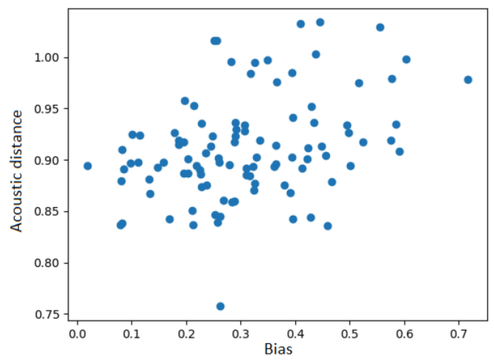


Figure 1: Scatter plot for the acoustic distance between wav2vec 2.0 embeddings against the bias for the SpSpecAug model on the HMI speech.

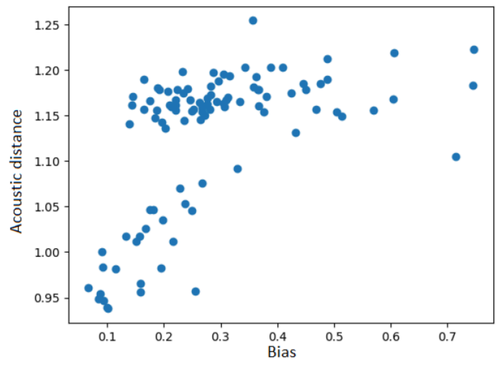


Figure 2: Scatter plot for the acoustic distance between wav2vec 2.0 embeddings against the bias for the SpSpecAug model on the read speech.

05 Conclusions

- The bias of the ASR system moderately correlates with the acoustics of the speaker.
- The acoustics quantification with the distances between wav2vec 2.0 acoustic embeddings reflected the bias more than XLSR distances.

06 Recommendations

Shortcoming: The bias-acoustics relationship may be nonlinear.
Recommendation: Use advanced statistical metrics to explore this nonlinear relationship.

Shortcoming: Noise impacts the quality of acoustic embeddings, as shown by differences in read and HMI speech.
Recommendation: Analyse isolated speech fragments (single sentences or words).

Shortcoming: Results may vary across languages; the study focused on Dutch.
Recommendation: Perform similar studies in different languages.

Shortcoming: The acoustic distances are not interpretable.
Recommendation: Check the relationship between the bias and isolated speech features like energy or pitch.